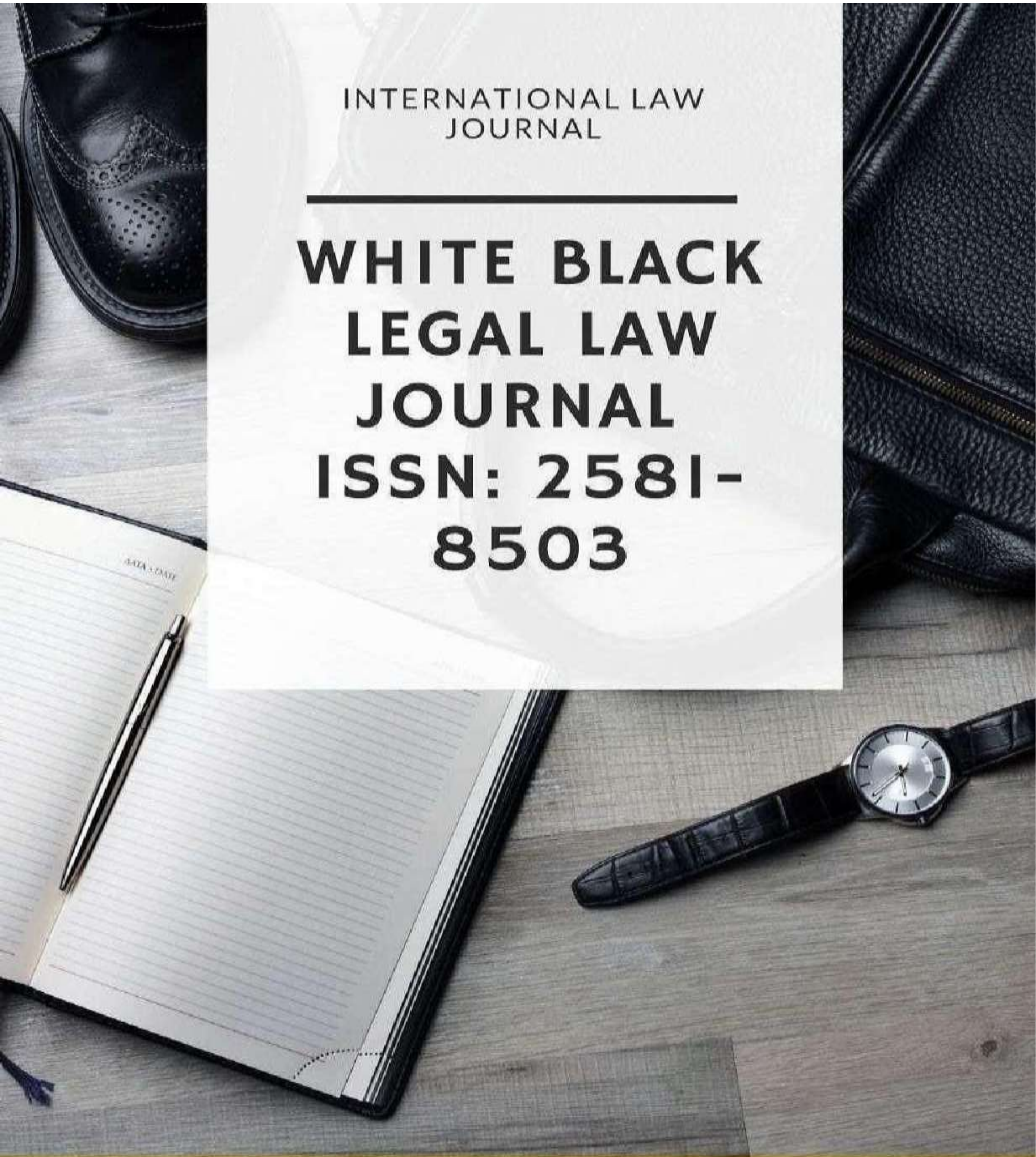

INTERNATIONAL LAW JOURNAL

WHITE BLACK LEGAL LAW JOURNAL

ISSN: 2581-8503

Peer - Reviewed & Refereed Journal

The Law Journal strives to provide a platform for discussion of International as well as National Developments in the Field of Law.

WWW.WHITEBLACKLEGAL.CO.IN

DISCLAIMER

No part of this publication may be reproduced or copied in any form by any means without prior written permission of Editor-in-chief of White Black Legal – The Law Journal. The Editorial Team of White Black Legal holds the copyright to all articles contributed to this publication. The views expressed in this publication are purely personal opinions of the authors and do not reflect the views of the Editorial Team of White Black Legal. Though all efforts are made to ensure the accuracy and correctness of the information published, White Black Legal shall not be responsible for any errors caused due to oversight or otherwise.

EDITORIAL TEAM

Raju Narayana Swamy (IAS) Indian Administrative Service officer

Dr. Raju Narayana Swamy popularly known as Kerala's Anti Corruption Crusader is the All India Topper of the 1991 batch of the IAS and is currently posted as Principal Secretary to the Government of Kerala . He has earned many accolades as he hit against the political-bureaucrat corruption nexus in India. Dr Swamy holds a B.Tech in Computer Science and Engineering from the IIT Madras and a Ph. D. in Cyber Law from Gujarat National Law University . He also has an LLM (Pro) (with specialization in IPR) as well as three PG Diplomas from the National Law University, Delhi- one in Urban Environmental Management and Law, another in Environmental Law and Policy and a third one in Tourism and Environmental Law. He also holds a post-graduate diploma in IPR from the National Law School, Bengaluru and a professional diploma in Public Procurement from the World Bank.

Dr. R. K. Upadhyay

Dr. R. K. Upadhyay is Registrar, University of Kota (Raj.), Dr Upadhyay obtained LLB , LLM degrees from Banaras Hindu University & Phd from university of Kota.He has succesfully completed UGC sponsored M.R.P for the work in the ares of the various prisoners reforms in the state of the Rajasthan.

Senior Editor

Dr. Neha Mishra

Dr. Neha Mishra is Associate Professor & Associate Dean (Scholarships) in Jindal Global Law School, OP Jindal Global University. She was awarded both her PhD degree and Associate Professor & Associate Dean M.A.; LL.B. (University of Delhi); LL.M.; Ph.D. (NLSIU, Bangalore) LLM from National Law School of India University, Bengaluru; she did her LL.B. from Faculty of Law, Delhi University as well as M.A. and B.A. from Hindu College and DCAC from DU respectively. Neha has been a Visiting Fellow, School of Social Work, Michigan State University, 2016 and invited speaker Panelist at Global Conference, Whitney R. Harris World Law Institute, Washington University in St.Louis, 2015.

Ms. Sumiti Ahuja

Ms. Sumiti Ahuja, Assistant Professor, Faculty of Law, University of Delhi,

Ms. Sumiti Ahuja completed her LL.M. from the Indian Law Institute with specialization in Criminal Law and Corporate Law, and has over nine years of teaching experience. She has done her LL.B. from the Faculty of Law, University of Delhi. She is currently pursuing Ph.D. in the area of Forensics and Law. Prior to joining the teaching profession, she has worked as Research Assistant for projects funded by different agencies of Govt. of India. She has developed various audio-video teaching modules under UGC e-PG Pathshala programme in the area of Criminology, under the aegis of an MHRD Project. Her areas of interest are Criminal Law, Law of Evidence, Interpretation of Statutes, and Clinical Legal Education.

Dr. Navtika Singh Nautiyal

Dr. Navtika Singh Nautiyal presently working as an Assistant Professor in School of law, Forensic Justice and Policy studies at National Forensic Sciences University, Gandhinagar, Gujarat. She has 9 years of Teaching and Research Experience. She has completed her Philosophy of Doctorate in ‘Intercountry adoption laws from Uttranchal University, Dehradun’ and LLM from Indian Law Institute, New Delhi.

Dr. Rinu Saraswat

Associate Professor at School of Law, Apex University, Jaipur,
M.A, LL.M, Ph.D,

Dr. Rinu have 5 yrs of teaching experience in renowned institutions like Jagannath University and Apex University. Participated in more than 20 national and international seminars and conferences and 5 workshops and training programmes.

Dr. Nitesh Saraswat

E.MBA, LL.M, Ph.D, PGDSAPM
Currently working as Assistant Professor at Law Centre II, Faculty of Law, University of Delhi. Dr. Nitesh have 14 years of Teaching, Administrative and research experience in Renowned Institutions like Amity University, Tata Institute of Social Sciences, Jai Narain Vyas University Jodhpur, Jagannath University and Nirma University.
More than 25 Publications in renowned National and International Journals and has authored a Text book on Cr.P.C and Juvenile Delinquency law.

Subhrajit Chanda

BBA. LL.B. (Hons.) (Amity University, Rajasthan); LL. M. (UPES, Dehradun) (Nottingham Trent University, UK); Ph.D. Candidate (G.D. Goenka University)

Subhrajit did his LL.M. in Sports Law, from Nottingham Trent University of United Kingdoms, with international scholarship provided by university; he has also completed another LL.M. in Energy Law from University of Petroleum and Energy Studies, India. He did his B.B.A.LL.B. (Hons.) focussing on International Trade Law.

ABOUT US

WHITE BLACK LEGAL is an open access, peer-reviewed and refereed journal providededicated to express views on topical legal issues, thereby generating a cross current of ideas on emerging matters. This platform shall also ignite the initiative and desire of young law students to contribute in the field of law. The erudite response of legal luminaries shall be solicited to enable readers to explore challenges that lie before law makers, lawyers and the society at large, in the event of the ever changing social, economic and technological scenario.

With this thought, we hereby present to you

SELF-REGULATION OF SOCIAL NETWORKING WEBSITES; NEW GOVERNORS OF ONLINE SPEECH, THEIR RULES, PROCESS AND APPROACHES

AUTHORED BY - BHAVNA RAJPUT
Rajputbhavna9gmail.com
Research Scholar, University of Delhi

The dawn of internet era revalorized the way we use to interact. Now with the emergence of SNWs we have blown up into new stage, now individuals are no longer passive recipient, but also are turning into publisher and broadcaster of his/her own speech. These SNWs facilitate participatory information sharing and collaboration in the creation of user generated content. However, great power comes with greater responsibility. The proliferation of SNWs has brought in its wake growing problems of illegal and harmful content online. The increasing misuse of social networking websites shoulder them to provide a safer platform to their users. This article provides an outlook of how algorithmic moderation system works. Further it examines some of the automated tools used by SNWs to handle child pornography, copyright infringement, nudity, terrorism and toxic speech. Further it identifies the ethical and other key concerns associated with it. Recent event suggests that due to sheer volume of data updated every minute, algorithm content moderation is the best method to deal with illegal and harmful content and it becomes inevitable to deal with increasing responsibility of platform, safety and security on global stage as this system remains inadequate, unaccountable and poorly understood . Despite the potential promise of algorithm this paper explores that even well optimized moderation system is also hampering right to free speech of users due to certain lacunas in its execution. Further paper concludes the major discrepancies of the algorithm mechanism such as, lack of transparency, unequal treatment, depolarisation andinabilityin understanding contextual clarity.

As we all know that we have entered in communication era named “web 2.0” where, from buying a needle to finding a soul mate, everyone can find anything online. Similarly, the introduction of social media to human being, makes people‘s life more convenient, democratic and speech enhancing.[1] Now they are able talk beyond the border in a very economic, affordable manner. These social networking websites turned everyone into publisher and broadcaster of their own speech.[2] The ease of use, affordability, and global richness unprecedentedly make it a global phenomenon which mankind has ever experienced.

Social networking websites, as they are famous for creating multichannel communication which provide platform to connect likeminded people with one another sitting in different corners of the world. Apart from the creation of informal friendship, SNWs have been used for other purposes i.e. to exhibit one's creativity, follow his passion, business promotion, business execution thus bringing people of every age group as a user. During the pandemic, social media has proved its significance and capabilities of bringing far sitting people closer and providing a widespread platform to disseminate news, information etc.

However, apart from this beautiful face, social media has a dark side too.**e**.g. shown by the Ask.fm case – a site where users have reportedly indirectly or directly caused nine teenagers to commit suicide.[3] SNWs received praise for the success of "Arab Spring" and many other movements but on the other hand incidents of suicide by a young girl due to bullying on ask fm, use of social networking website in Christchurch incident, Neo-Nazi, Rohingya massacre reveal the other dangerous side of SNWs.

Various reports[4]by international and national institution revealed that crime through social media is rising each day. These crimes include person centric crimes-such as bullying, cyber stalking, cyber harassment, impersonation, and content centric crime, such as availability of pornographic &obscene content, hate speech, islamophobic content, drug abuse, child pornographic content. Content centric

[1] Louis W. Tompros, Richard A. Crudo, Alexis Pfeiffer, The Constitutionality of Criminalizing false Speech Made on Social Networking Sites In A Postalvarez, Social Media-Obsessed World,31, HJLT,65, 66,2017

[2]Human Rights Council, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, 17th Session, A/HRC/17/27 (May 16, 2011), available at: http://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf

[3]Richard Steppe,The freedom of speech on social networking services Do we need protection against our own expressions,Jura Falconis, Jg. 50, 2014-2015, nummer 3, 559,559,

[4]2016, Child Online Protection in India: Unicef India,House of Commons Home Affairs Committee Hate Crime: Abuse, Hateand Extremism Online Fourteenth Report Of Session 2016–17

crimes pose greater challenge before the government and SNWs itself, as they affect the society at large and sometimes are capable enough to disturb the social harmony of a nation. For example, circulation of a few minutes video resulted in Muzaffarnagar riots. Similarly, a simple tweet by a techie sitting in Bangalore caused north eastern exodus in Maharashtra. Along with these few, numerous incidents have been reported when a simple tweet or post led to unthinkable loss.

A study[5] reveals that terrorist and extremist groups have been using SNWs i.e. facebook, twitter etc. to advocate their ideologies, recruitment, incitement, spread radical thinking. Similarly, hate speech, gender and racial discrimination speech and holocaust denial expression offend wide-held public values which may incite violence and can disturb public harmony. Likewise, the report of UNESCO on the accessibility of child sexual abuse imaginary through SNWs has raised concern regarding the content regulation across the globe.

The increasing misuse of SNWs in last few decades has led government and private entities to impose greater regulation on the SNWs through the use of strict regulation, to deploy increasingly sophisticated technological mechanism to block, remove, monitor and adopt restrictive legislation to justify it[6]. Almost all the nations are striving to find ways of regulation of SNWs through introduction of new laws or through the amendment of already existing laws.

Content regulation on SNWS has always been subject of debate. It has been contested that regulation of SNWs is not feasible on domestic level, as it is terra nullius,[7] it might raise cross jurisdictional and policy issues as well. One another argument against the state regulation is that state authorised regulatory process is too much time taking, unable to provide speedy recovery which is true to an extent. Whereas, argument in favour of self- regulation is that the architecture of internet is configured in a manner that it is better regulated by self regulation.[8]

In order to avoid the shortcoming of state based regulation, and provide a safer platform to its user

[5] United Nations Office on Drugs and Crime Vienna, Handbook on Children Recruited and Exploited by Terrorist and Violent Extremist Groups: The Role of the Justice System

[6] A/HRC/17/27

[7] Iibid,

[8] Yar, Majid (2018) "A Failure to Regulate? The Demands and Dilemmas of Tackling Illegal Content and Behaviour on Social Media," International Journal of Cybersecurity Intelligence & Cybercrime: 1(1), 5-20.,9

and protect their vulnerable users for some type of specific content, and eliminate the liabilities arising due to court cases online hosts itself regulate the content. Moreover, the architecture of Internet makes self-regulation mechanism more feasible and it may exemplify the shift from government to governance[9]. Beyond the provision it sets out that the regulation of online content has been controlled by ICHs, for example, Facebook make as a precondition of use the adherence to predefines community standard covering issues such as obscenity, child pornography, nudity, sexualised image, threats bullying, procurement &sale of prohibited goods and sexual services. Similarly, Twitter's user agreement prohibited behaviours including IP violation, posting abusive and hateful content, unauthorised sharing of someone else's private information or intimate images, distribution of spam and malware. Other popular social networking websites such as YouTube, Instagram, Snapchat also institute compatible community standards upon users.

In 2016 during American presidential election, the role of social networking platform and the information they circulate online has been questioned by a public concern, for the first time in significant numbers, about the way the content is produced. In 2017 self -regulation of content becomes a hot topic after a series of highly publicised violent tragic events were broadcasted in some cases live to the world via Facebook or others social networking platforms. These events raised questions in public domain about what and how material circulates online,how it is regulated what is the regulatory mechanism behind the regulatory tools and how these companies ensure that legitimate speech will not be hampered in course of regulatory process.

Indian Legal Framework –

Rule 3 of the intermediary guideline[10] empowers the intermediary to regulate content on their own. First rule 3(1)(b) provides a list of content that needs to be regulated on SNWs. 3(1)(d) deals with the removal or restriction of access to allege content by the court order or order of a competent authority. Whilst ruled 3(2) provides power of self-regulation under the head of grievance redressal mechanism of intermediary

It provides that intermediary shall appoint a grievance officer and shall publish its name and number on the website or mobile based application. Further, it provides that grievance officer shall

[9] ibid

[10] The Information Technology (Intermediary Guidelines And Digital Media Ethics Code) Rules, 2021

acknowledge all the complaints within 24 hours and dispose them off within 15 days from the date of receipt of complaint. Further it provides that some content hasto be dealt on priority basis and should be disposed off within 24 hours from the receipt of complaint made by individual or any person in his behalf with respect to any content by which he is personally aggrieved.

Types of Content Moderation and Stages of Content Moderation-

Commercial content moderation is the systematic practice of reviewing user-generated content posted on internet sites, SNWs, and other online outlets.[11] The activity of reviewing user-generated content may take place either before or after the publication of content on the website. The prior screening of user generated content known as ex-ante and post publishing screening known as ex-post content moderation.[12] In both processesof removing illegal and harmful content, the decision can be made manually by human moderator or automatically through algorithmic machine learning software. However, former procedure was predominately based on algorithm software and In later process, moderators passively work to remove content and content screening may be triggered due to complaint received against specific content from third party (for example, companies alleging misappropriation of material they own), or from other users who are disturbed or concerned by what they have seen.

3.1 Ex-ante Content Moderation-This process of moderation takes place between the submission or inclusion of particular data and publication thereof.[13] When a user uploads any media (picture, video, audio) on his wall profile over the social networking website, generally a message box pop up on top of the screen; "your post is being processed, we will shortly notify you, once it's done". No longer, another message pops up on the screen that "your post is ready to view".[14] Ex- ante content moderation is the process which takes place in between two messages, upload of post and publication of post. Most of the time, ex-ante moderation is an automatic process largely based on algorithm screening without active interference of human decision making.[15] This process is generally known as filtering, in this process software automatically restrict access to a particular content containing

[11] The Virtues Of Moderation James Grimmelmann! 17 Yale J.L. & Tech. 42 (2015)

[12] The New Governors: The People, Rules, and Processes Governing Online Speech Kate Klonick,1599,1618

[13] Ibid 1638

[14] Video facebook: help centre

[15] Louis W. Tompros, Richard A. Crudo, Alexis Pfeiffer, The Constitutionality of Criminalizing false Speech Made on Social Networking Sites In A Postalvarez, Social Media-Obsessed World,31, HJLT,65, 66,2017

specific keyword(in case of text) i.e.child pornography. This filtering mechanism also works with visuals, imageswhich depict blood, extreme cruelty, brutality and violence, gender sensitivecontent etc.

An example of content that can be moderated by these methods is child pornography, which can reliably be identified upon upload through a picture-recognition algorithm called PhotoDNA.[16] Geo-blocking is another type of automatic ex-ante moderation. Unlike photo DNA which prevents the publication of illegal, harmful and inappropriate images, geo-blocking prevents both publication and accessing of alleging content based on users' location. Geo-blocking is very crucial with respectively to the material which has been declared offensive in any particular country on territorial basis. As happened in controversy over the *Innocence of Muslim videos*, geo-blocking generally proposes at the request of a government notifying a platform that a certain type of posted content is illegal or harmful according to its local laws.[17]

Although algorithms itself cannot automatically decide what kind of content should be blocked from being published. Content scrutinised automatically by algorithm is typically a content that can reliably be identified by software as illegal or otherwise prohibited as per the community standard of the platform. The universe of content that is automatically moderated ex-ante is regularly evaluated and updated through iterative software updates and machine learning.

3.2- Ex-post Content Moderation- Unlike ex-ante content moderation, ex post content moderation comes into the picture only after the publication of content on platform. As previously discussed, ex-ante method is not 100% accurate and has its own limitations. As an estimate 800 hour content has been upload every hour on social media, if we took an estimate of 10% of error that there is very much offensive and harmful content remained on social networking websites. Therefore, SNWs also use post publication screening where content is scrutinised by the moderators after the publication either suo-moto or upon getting notified by users. It can either be pro-active or re-active.

3.2.1 EX- Post Pro-active manual content model

As the name suggests name, ex post moderation can be initiated only after the publication of the

[16] Tracy Ith, Microsoft's PhotoDNA: Protecting Children and Businesses in the Cloud, MICROSOFT: NEWS (July 15, 2015), https://news.microsoft.com/features/microsofts-photodnaprotecting-chi

[17]Kyle Langvard, Regulating Online Content Moderation,The Georgetown Law Journal,Vol. 106:1353

content. However it might be proactive and reactive. It is true that it has been initiated after the publication but not necessarily upon receiving complain against the content. It is known as pro-active because it is initiated automatically by machine learning without getting any notification from the human moderators or the users' report. It's a repeating process of screening content which has already scrutinised through ex-ante moderation to ensure more accuracy and safety of users. It removes the already published material which is inconsistent with platform's policies. Almost all the platforms are engaging in proactively seeking out and removing published alleged content. Currently, this proactive method is being used for removing violent, terrorist, extremist speech. As of 2016, Facebook has proactively removed 90% post and profiles linked to terrorist, extremist activity[18]. This is an important development affecting content moderation, which seeks to strike an ever evolving balance between completing interests, ensuring national security and maintaining individual liberty and freedom of expression.

3.1.2 Ex-post Reactive Content Moderation-

Apart from the exception of pro-active moderation for terrorist extremist content explained above, almost all user generated content which has been published on websites is reviewed reactively, predominantly through flagging/ reporting by the users of websites. The flagged content is reviewed by human moderators against community standards and internal moderation guidelines of platform.

Algorithm Method for Moderation-

Algorithm method is important but still under examined method of the increasingly evolving content moderation techniques grouped under the generic term of artificial intelligent (AI).[19] Recognising the significant technical advancement of machine learning, automated tools are not only being increasingly used to fulfil crucial moderation function but also actively heralded as the force that will somehow save moderation form its existential challenge.[20] While defending self-serving and unrealistic narrative about their technological powers Facebook CEO, Mark Zuckerberg during his congressional testimony in 2018, notably accepted that AI is the future solution to Facebook's current

[18] UNOCT report 2021, COUNTERING TERRORISM ONLINE WITH ARTIFICIAL INTELLIGENCE: An Overview for Law Enforcement and Counter-Terrorism Agencies in South Asia and South-East Asia

[19] Robert Gorwa, Reuben Binns and Christian Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, Big data and Society, Sage Journal

[20] ibid

political problems.[21] It is not only a hypothetical hype, the statistics revealed through companies transparency reports and media reports have shown that automation play a significant role in content moderation[22] i.e. after a public controversy, Facebook improved its “Myammar language hate speech classifier”, leading to 39% hike in take-down form automation flags in only six months; YouTube’s new reports show that 98% of videos removed under “violent extremists content” are flagged by machine learning algorithm, and Twitter recently affirmed that it has taken down hundreds of accounts that tried to spread terrorist propaganda, with some 93% consisting of account flagged by “internal property Spam fighting tools”.[23]

Algorithm Method- What is this and How It is used

Automated system came to existence when manual moderation system became unfeasible or less effective due to increasing traffic on social networking website. ‘In short, Algorithm content moderation has been defined as a system that classify user-generated content on the basis of either on matching or prediction, and leading to a decision and governance outcome i.e. removal, geo-blocking account suspension.[24]

Algorithm method in practice by SNWs

6.1 Child sexual abuse- Photo DNA is the majorly adopted technology by SNWs to identify child sexual abuse. It was initially developed by Microsoft Research and further modified by Henry Fared to help National Centre of Missing and Exploited Children (here in after NCMEC) with the aim to filter hidden child pornographic content. Further, it is used by the Microsoft’s own service Bing and One Drive along with Google, Twitter, and Facebook, Adobe system, reddit and NCMEC. Later on Microsoft donated the Photo DNA technology to project VIC by NCMEC.

Photo DNA works on a technology called “robust hashing” which calculates the particular characteristics of a digital image, this “digital fingerprints or hash value” is used to match it with other

[21] The Guardian, The key moments from Mark Zuckerberg's testimony to Congress,https://www.theguardian.com/technology/2018/apr/11/mark-zuckerbergs-testimony-to-congress-the-key-moments

[22] Macdonald S, Correia SG, Watkin A-L. Regulating terrorist content on social media: automation and the rule of law. *International Journal of Law in Context*. 2019;15(2):183-197. doi:10.1017/S1744552319000119

[23] GIFT 2019

[24] The Virtues Of Moderation James Grimmelmann! 17 Yale J.L. & Tech. 42 (2015)

copies of that same image. Photo DNA provides a way to create a unique signature - similar to a fingerprint - forming a photo that will remain consistent even after the images are edited or manipulated, the photo DNA is computed by converting it into black and white, resizing it, and breaking into grid. In each grid, cell a histogram of intensity gradient or edges are found. This photo DNA is created from this gradient information. While you cannot reconstruct the photo from its DNA, if two images have similar DNA, software knows that they are the same. It is more efficient and effective if amount of data in the photo DNA is small, it help us in quickly finding the matches, across large data sets which allow finding the needle in the haystack.

6.2-Terrorism- Due to increasing use social networking platform by terrorist to commit terrorist activity, the European Commissioner announced the establishment of the European Union Internet Forum in Dec 2017 which brought European Union official together with the representative of four major Internet online industries including Facebook, Google, Microsoft and Twitter.[25] After 6 months of announcement and two meetings, the members of Internet forum announced the EU Code of Conduct on Counter monitoring Illegal and hate Speech Online.[26] Imposing a wide range of principles including, removal of hate speech within 24 hours of notification under the term of services, and to encourage co-operation among themselves and other online platforms to enhance sharing of best practise, knowledge and research to fulfil the commitment for Internet forum established the GIFCT in 2017, an organisation which remains highly secretive and publishes a little about it operations.[27] The main object of the organisation was improvement of efficient automated system for removing extremist images, text and videos.

6.2 Toxic speech- It is used as umbrella term for various concepts including bullying defamation claims, harassment, hate speech, profanity, personal attacks etc. toxic speech is regulated by curating a large corpus of text, manually labelled as toxic then automatically classified to flag it.

In 2017, Jigsaw, a Google subsidiary focused on global security challenges announcing a new project named “perspective API” (an application programming interface) with a stated aim to make it easier

[25] Gorwa R (2019a) The platform governance triangle: Conceptualising The Informal Regulation of Online Content. Internet Policy Review 8(2).

[26] Fiedler K (2016), EU Internet Forum against Terrorist Content and Hate Speech Online: Document pool.

[27] GIFCT (2019) About the global internet forum to counter terrorism

to host better conservation. As per the description of project regarding the use of "perspective API", platform could receive a score which predict the toxicity of text, which could be used to give feedback to commenter or helping moderators in moderation.[28]Similar efforts have been done by other SNW including Twitter and Disque (a third party comment penguin provider).[29] In recent past, Facebook has been pressurised by EU members to combat hate speech and it responded by developing classifier trained to predict toxicity of underlying text as per the score, and then automatically flagged it for humans review. Initially this effort was limited to certain languages including English and Portuguese and now it has been extended to other languages i.e. Burmese.[30] Instagram has also developed toxic speech classifier to identity comment for bullying, harassment, adopting a different approach from Facebook by offering an Opt- out filter where user can hide the comments rather than referring it for moderation. You Tube also moderates toxic speech by machine learning classifiers that seek to predict the incident of harassment, hate speech, vulgar, swearing and inappropriate language in a video, in order to demonetise it and also barring advertisers to associate their brand with such content that could ruin their goodwill.[31].

6.3 Copyright-

Copyright has historically been one of the domains where the strong economic interests demand technology to protect their work from infringement following the Napster and file sharing controversies[32]. In the early 2000s it coincided with the risk of the social networking platform to boost and share creative artistic cultural work in video sharingplatform mainly as they become a key target for companies in other rights stakeholders seeking to licence distribution of their content on the platform. After the YouTube's acquisition by Google in 2006, Viacom sued platform for its copyright infringement by YouTube users. This case was finally settled in 2013,this long running litigation and fear of compensation increased pressure on the platform to monitor and police the content more strictly on their platforms even before getting notified.[33]

[28]Robert Gorwa, Reuben Binns and Christian Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, Big data and Society, Sage Journal

[29] ibid

[30]ibid

[31] Internet Creators Guild (2016) YouTube de-monetization explained;**https://medium.com/internet-creators-guild/youtube-de-monetization-explained-44464f902a22**

[32] Robert Gorwa, Reuben Binns and Christian Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance, Big data and Society, Sage Journal

[33]ibid

Anticipating the growing economic and political pressure, in 2006, YouTube started to explore the best practices of content moderation that were formally and procedurally independent from notice and take down procedure and can be run simultaneously. In 2006, the experimental effort came out as Content ID system that YouTube has been deploying for more than a decade. Like other SNWs, YouTube has remained secretive about the technical implementation of its propriety algorithmic moderation.[34] Nonetheless some characteristics can be visible based on publically available material. ContentID algorithm does not only find identically similar files but also identifies different performance of cultural work that may be protected under copyright laws. This mechanism not only finds multiple upload of a music video, but also recording of performance, edited performance of that song, audio etc. Through perceptual hashing, the resulting fingerprints aim to reflect characteristics of the audio or video content: each note in a song could be represented by presence or absence of specific frequency values; the volume of a particular note, by some amplitude at frequency corresponding to musical note.[35]

The main challenge before content ID is over blocking. However, content ID and other automated system may improve from a technical standpoint enhancing their ability to create quality fingerprints and then accurately detect those fingerprints but it does not necessarily mean that they become more adept at evaluating actual copyright infringement.

In some cases, Copyright law allows the exception of third party fair use which varies across jurisdiction but create important exemption for educational purposes, parody and few other context. Permission of fair use standard cannot be programmed in ML system, thus manual and institutional oversight is essential. Path of copyright protection paved by YouTube is followed by Facebook and Instagram by deploying Right Management which works on similar functionality to content ID

6.4 Obscene Nudity Detectors for Two DNA

Nudity detection API- It works on the basis of data sets which contain nude and non- nude photos

[34] In addition to the officially published material, a leaked ‘YouTube Content ID Handbook’ is circulating online that had apparently been prepared by YouTube for rightsholders. See the last available version, updated Q2/2014 at https://scribd.com/document/351431229/YouTubeContent-ID-Handbook.

[35] Duarte N, Llanso E and Loup A (2017) Mixed Messages? The Limits of Automated Social Media Content Analysis. Washington, DC: Center for Democracy & Technology. Available at: https://perma.cc/NC9B-HYKX

crawled from different online platforms, sites. It crawled around 2000000 nude images from different nude picture websites and forum and non- nude pictures were sourced from Wikipedia. It works on the basis of classifying algorithm. These data sets were randomly split into a train (80%), validation (10%) and the test set (10%). The accuracy of classifier depend/ trained on train set comes out to be slightly over 95%.

Human behind the Moderation

7.1 Moderators – As previously stated that algorithm contentmoderation became inevitable due to increasing traffic but still human moderation is a crucial part of content moderation process and in all the moderation practise human moderators should be on loop to check and verify the accuracy of automatic moderation. Human moderators are the essential part of moderation chain.

When flagged content is sent to server for its review by a human moderator, mostly, there are three tiers of content moderation on almost all SNWs. T-3 moderators generally deal with day to day moderation, almost all flagged content is reviewed by them. T-3 moderators specifically review the content that has been reported in categories of lower priority such as nudity, defamatory content, pornography, inappropriate or annoying content that is humiliating, insulting or attacking based on religious affiliation and sexual orientation, content that promotes violence to a person or animal. T-2 moderators generally work in capacity of superiors of tier 3 moderators and they randomly cross check the review decision took by T-3 moderators. Along with this T 2 moderators review prioritise content such as child pornographic content, child sexual abuse content, immediate threat of violence, suicidal, self harm content and terrorist, extremistcontent. T-1 moderators are mainly lawyers or policymakers based at headquarter[36]. Mostly SNWs do content moderation in two ways; either they directly hire content moderation team or outsource much of their content moderation work to third party companies indulge in this business i.e. Sutherland deloitee, O-desk (now up work).[37] In 2009, Facebook opened an office in Dublin, Ireland, that had twenty dedicated support and user-operations

[36] Telephone Interview with J.L., Tier 2 Moderator, Facebook (Mar. 11, 2016). J.L. was a Tier 2 moderator based in the Eastern United States

[37] Telephone Interview with Sasha Rosse, supra note 29; Adrian Chen, Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where "Camel Toes" Are More Offensive than "Crushed Heads," GAWKER), http://gawker.com/5885714/inside-facebooks-outsourced-antiporn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads [https://perma.cc/ HU7H-972C]; Chen,

staff.[38] In 2010,

YouTube started a new initiative launching in 2016 “the heroes programme” whichdeputies users of the platform in actual participation in content moderation process in exchange of some perks and priorities such as access to exclusive workshop, fan fest and sneak preview product launches.

Conclusion-

The use of automatic mechanism helps to remove illegal and harmful content more quickly and effectively. ICHs are trying to evolve move efficient and error free mechanism to improve their precision. But it is also not free from loopholes. It is being argued that this method is not democratic thus hampering the freedom of speech and expression provided under Art 19 of constitution of India. One of the critical analysis about algorithm moderation system often emphases on technical challenges; that these system predominantly are facing now and will face in future as well .One of the major concerns is over blocking. It is very commonly argued that it is difficult for ML to take complex contextual decision on slippery concept like toxic speech; hate speech etc. and automation algorithm system make thousands of incorrect decisions on daily basis. Mainly theproblems jointly pointed out by civil societies, lawyers and various reports include fairness, transparency and accountability in machine learning. As far as accuracy of the new machine learning system is concern, machine learning is subject to attack in adversarial scenarios. One type of the vulnerabilities of ML algorithm is that adversary can change the algorithm prediction score by slight change in input, often getting unnoticed by human.

[38] Telephone Interview with Sasha Rosse, supra note 29